Creating and exploiting multimodal annotated corpora

Philippe Blache[(1)], Roxane Bertrand[(1)] & Gaëlle Ferré[(2)]

(1) *Laboratoire Parole & Langage, CNRS & Aix-Marseille Universités*
29, av. Robert Schuman. 13100 Aix en Provence

(2) *LLING - Université de Nantes*
Chemin de la Censive du Tertre. BP 81227. 44312 Nantes cedex 3
e-mail: {blache; roxane.bertrand}@lpl-aix.fr, gaelle.ferre@univ-nantes.fr

Abstract

The paper presents a project of the *Laboratoire Parole et Langage* which aims at collecting, annotating and exploiting a corpus of spoken French in a multimodal perspective. The project directly meets the present needs in linguistics where a growing number of researchers become aware of the fact that a theory of communication which aims at describing real interactions should take into account the complexity of these interactions. However, in order to take into account such a complexity, linguists should have access to spoken corpora annotated in different fields. The paper presents the annotation schemes used in phonetics, morphology and syntax, prosody, gestuality at the LPL together with the type of linguistic description made from the annotations seen in two examples.

1. Introduction

In recent years, linguists have become aware that a theory of communication describing real interactions should involve a complexity of dimensions which is why linguistics and NLP have turned to multimodal data where the complexity of speech is better represented. Each dimension is itself composed of a set of diverging parameters and must then be related to the other dimensions of speech. However, annotating such inputs remains problematic both for theoretical and technical reasons. First, we still need a linguistic theory taking into account all the different aspects of multimodality, explaining in particular how the different linguistic domains interact. At the same time, we need to specify a standardized way of representing multimodal information in order to give access to large multimodal corpora, as richly annotated as possible. What is meant by large corpora is however quite a relative notion since in some linguistic fields such as syntax for instance, corpora of several million words are used whereas in prosody where most of the annotations are made manually, a few hours of speech are considered as a large corpus.

This paper describes the first results of a project aiming at answering these different issues. In the first section, we specify a coding scheme adapted for multimodal transcription and annotations. In a second part, we describe the automation of the production of multimodal resources by means of a platform integrating different annotation tools. This platform consists in a sequence of tools leading from raw data to enriched annotations relative to each linguistic domain. We illustrate the application of this environment by the description of a large multimodal annotated corpus for French. Finally, we briefly sketch some first results obtained thanks to this resource.

2. A multimodal coding scheme

Coding schemes for multimodal annotation have been developed in several projects such as MATE, NIMM, EMMA, XCES, TUSNELDA, etc. What comes out of the various coding schemes is mainly that they are very precise in one or two modalities. However, they generally do not cover the entire multimodal domain nor the very fine-grained level of annotation required in every modality. We propose to combine the existing schemes and to extend them so as to obtain an XML coding scheme that would be as complete as possible in all the following domains:

- Corpus metadata: we will use a TUSNELDA-like coding scheme ([Tusnelda05]) in which all the information such as speaker name, sex, region, etc. is noted.
- Morphology and Syntax: we propose to adapt the Maptask coding scheme for the French language in the morphological dimension, completed with syntactic relations and properties.
- Phonetics and prosody: some annotations are inspired by MATE ([Carletta99]), and are completed with other type of information. The phonetic representation is coded in SAMPA. Prosodically, we adopt the coding scheme proposed in Di Cristo et al. (2004) in which the main prosodic information is annotated in a manual and automatic way: we use the INTSINT and MOMEL algorithms as a first step to represent the phonological level of intonation.
- Gesture analysis: we adapt the MUMIN coding scheme ([Allwood05]) yet coding separately gestures and discourse tags.
- Pragmatics and discourse analysis: we use the Maptask ([Isard01]) and DAMSL coding schemes, extended to other discourse types such as narration, description, etc.

As for gestures, the coding scheme concerning more specifically facial expressions and head movements is based on the FACS standards, based on different proposals ([Kendon04], [Kipp04]). The gesture typology is encoded following the scheme proposed in [McNeill05]. A gesture lexicon is compiled from the existing descriptions found in the literature ([Kipp04], [Krenn04])

and on the basis of our own experience. The following descriptions illustrate some annotation conventions at different levels:

Morphosyntax

```
Token:: attributes: orthography
        content: Lex*
```

```
Lex::   attributes: id category lemma rank
probabiblity frequency phonemics reference
        content: msd

   category:  {Adjective  Determiner  Noun
Pronoun Adverb Preposition Auxiliary
   Verb  Conjunction  Interjection  Ignored
Punctuation Particle Filled pause}
```

Gestures

```
Head::
   attributes:Movement_Type        Frequency
Horizontal_Plane Vertical_Plane Side_Type
   Movement_Type: {Nod, Jerk , Tilt , Turn ,
Shake , Waggle , Other}
   Frequency: {Single , Repeated }
   Horizontal_Plane: {Forwards , Backwards ,
Sideways}
   Vertical_Plane: {Up, Down}
   Sid_Type: {Left , Right}
```

3. The annotation platform

Until now, corpus annotation was essentially based on written corpora, the annotation of oral corpora being very limited. Some transcribed oral corpora exist, but they rarely contain other information domains such as phonetics and prosody.

The problem comes from the lack of tools or more precisely, the difficulty in integrating them with a common format. We have developed a platform addressing these questions. It provides help at each step of the process, from raw data to high-level annotations.

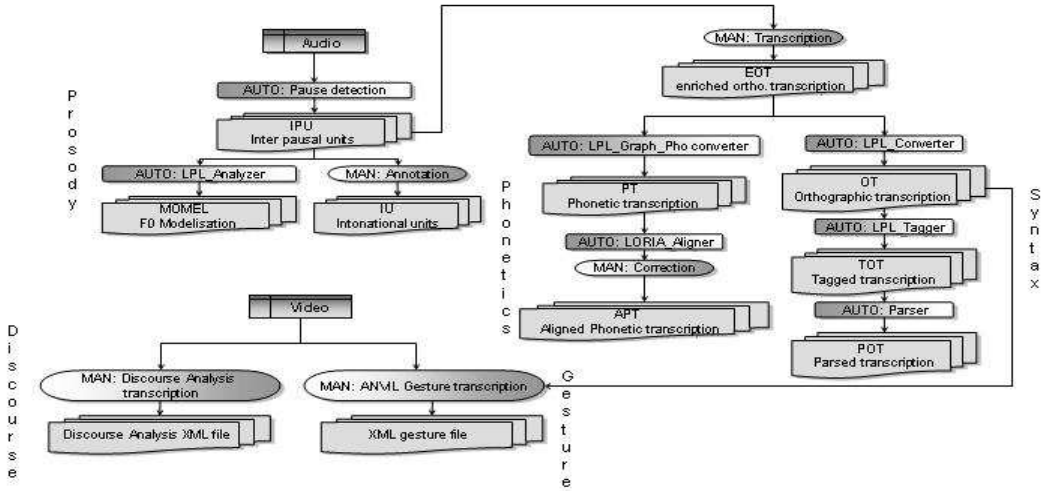

Figure 1: *Multimodal annotation process*

Figure 1 describes the general process in which the automatic (AUTO) or manual (MAN) treatment at each step is specified.

- **Segmentation in Interpausal-Units:** Transcriptions are done starting from an automatic pre-segmentation of the speech signal into interpausal-units (IPU) that are blocks of speech bounded by silent pauses of at least 200 ms. IPU segmentation facilitates the transcription, the phonetization and the alignment with the signal. Moreover, speech overlap phases were extracted from IPU.

- **Transcription:** Transcription conventions are derived from [Blanche-Benveniste87] on top of which other information is added (such as elisions, particular phonetic realizations, filled pauses, false starts, repetitions, truncated words, etc.). From this initial enriched orthographic transcription, two transcriptions are derived: one is phonological, the other is phonetic. The following example illustrates this step:

- *EOT*: `et c(e) qu(i) était encore plus le choc c'est que en [fait, faiteu]`
- *Phonologic version*: `et ce qui était encore plus le choc c'est que en fait`
- *Pseudo-phonetic version*: `et c' qu était encore plus le choc c'est que en faiteu`

- **Phonetization:** This step produces the list of phonemes. After a tokenization, the symbolic phonetizer (see [DiCristo01]) provides a list of tokens and their phonetization labeled in SAMPA. The TOE may sometimes be difficult to use, and a direct phonetic transcription can be, in some cases, simpler for the transcriber; the phonetizer therefore accepts mixed orthographic and SAMPA symbols as an input; SAMPA symbols come out unchanged in the output. The TOE is consequently more precise and standard.

- **Alignment:** The aligner takes as input the list of the phonemes and the audio signal. It then localizes each phoneme in the signal.

- **Prosody :** Prosodic annotations essentially encode the prosodic categories (intonation and accentual units) and the intonation patterns associated to them. Such annotations are done by experts exclusively. We also use the INTSINT system (see [Hirst and al. 00]) which does not suppose any a priori knowledge of the phonological system of the language. The interest to have both manual annotations and automatic INTSINT annotations is to improve INTSINT itself, but also the knowledge, which is still very fragmentary, of the prosodic domains in French. This coding is achieved automatically and is naturally integrated in our system. Basic data are temporal series (audio recording, physiological, aerodynamic parameters, etc.). Until these last years, most of these signal corpora were annotated by tabulated label files: a label is a time value associated to a list of attributes and values. These annotations are made by hand (most frequently) and/or automatically.

- **Morphosyntax:** morphosyntactic annotation, is done automatically, using a POS–tagger (LPLsuite, cf. [VanRullen05]) that has been adapted to spoken language. The system has been trained with appropriate data, and custom correcting code heuristics has been developed. It is then checked and corrected manually.

- **Syntax**: We have developed an original statistical parser, adapted for the treatment of spoken data. This has been done in two different phases. The first consisted in parsing a spoken language corpus by means of a symbolic parser (cf. [Blache05]). In a second stage, the output, has been corrected manually, the result being a treebank for spoken language. Finally, the statistical parser has been trained on these data. The tool we obtain is used in order to generate automatically the trees of the corpora to be annotated. This output also has to be checked manually.

4. A first annotated corpus

The Corpus of Interactional Data (CID, http://crdo.up.univ-aix.fr/) is an audio-video recording of spontaneous spoken French (8 hours, 8 pairs of speakers). If features data recorded in anechoic room and containing 110.000 words. Each speaker of the dialogue is equipped with a headset microphone enabling the recording of the two speakers' voices on two different tracks. This enables the study of speech at a phonemic and prosodic level. It also enables the study of overlapping speech which is frequent in spontaneous interactions but seldom analyzed because of the difficulty to separate the intertwined voices of the two speakers a posteriori. Yet, overlapping speech plays a main role in conversation and requires experimental investigations.

In this corpus, we aimed at collecting useful data for the investigation of all the levels under study: the audio quality is optimum and they have been videotaped with a high quality digital camera. The corpus, described in [Bertrand07a], has been annotated following the different steps described above.

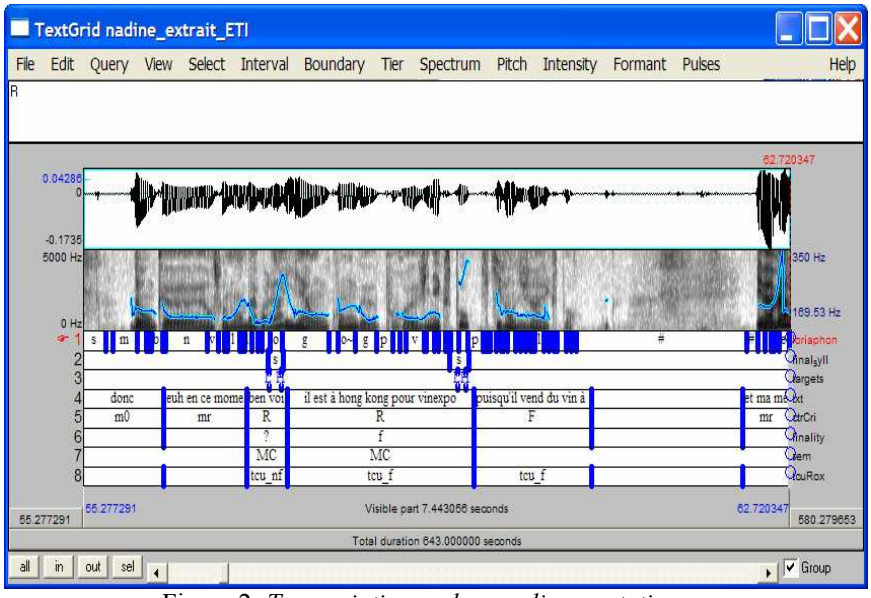

Figure 2: *Transcription and prosodic annotations*

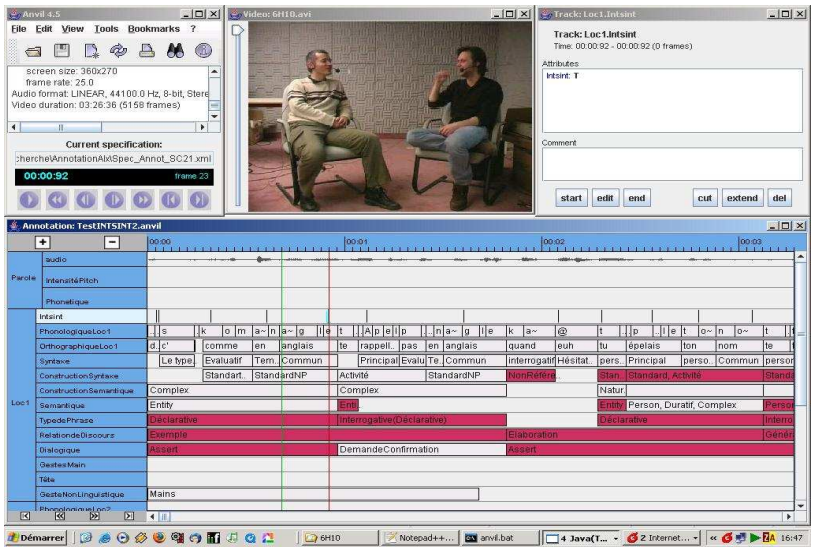

Figure 3: *CID annotation board*

We then aligned the orthographic and phonetic transcription with the signal and added information from different linguistic fields (prosodic units, pitch contours, morphosyntactic categories, syntactic phrases), as illustrated in Figure 2. The tiers bear different information, from the phonetic segmentation (tier 1) to the conversational units (tier 8). This figure illustrates the different pitch contours in their formal and functional aspects (tier 5 and 7). These annotations have been done separately on the audio files and constitute the basis of our present project which consists in the annotation and processing of the corpus from a multimodal perspective.

The annotation of the gestures made by the participants is being done manually using ANVIL as shown in Figure 3.

5. Some results

5.1 Gestural backchannels

Backchannels are signals produced by the hearer (the co-participant) during a dialogue (*mmmh*, *yes*, *ok*, etc.), *cf.* [Bertrand07b] They have different functions such as acknowledgement, assessment, acting the participation to the dialogue, etc. If vocal backchannels are very frequent, gestural ones (head movements, smiles, eyebrow, etc.) also play an important role. The question is to see whether vocal and gestural BCs behave similarly. The following example, taken from the CID corpus, illustrates such phenomena, vocal BCs being represented in italics, and gestural ones in frames.

```
A ah ouais nous on est rentré à (...) dix heures
dix heures et demi je crois du soir (...)
B                                  nod
A et elle a accouché à six heures je crois (...)
B                       ah quand même ouais
B                       head tilt / eyebrow raising
A donc c'était ouais c'était quand même assez long
quoi (...)
B head tilt
```

```
[A] oh yeah we were admitted at 10, 10.30 I think
pm
[A] and she had the baby at 6 I think
[B] [oh yeah right?]
[A] so it was yeah it was quite long indeed
```

What we observe is that vocal and gestural BCs show similar behavior, in particular concerning the morphological and discursive production context. They appear after nouns, verbs and adverbs, but not after connectors or linking words between two conversational units. As for prosody, gestural BCs can occur after accentual phrases and intonational phrases (IP) whereas vocal BCs only occur after IPs. Both BCs seem to be favored by rising and flat contours. In conclusion, we can say that vocal and gestural BCs occur in the same kind of environment but gestural BCs seem to be delayed as compared to vocal ones (they occur after the end of the intonational unit). Gestural BCs are also encouraged when the speaker is gazing at the interlocutor. Moreover, BCs are produced after some completion point but not often in places of possible turn change.

5.2 Reinforcing gestures

Some gestures can be produced by the speaker without being prompted by the listener (with phatic eyebrow movements, gaze direction, head movements, etc.), *cf.* [Ferré07] Such gestures are intended to reinforce discourse, and play a specific role in the communication process.

```
          elle était SUper stricte elle voulait PAS...
tu vois elle interdisait que tu sortes
head      nod        shake
hands               beat
gaze                gazes at interlocutor
```

```
[A] she [the teacher] was super strict she didn't
want... you see she forbade us to leave
        the room [during lessons]
```

From the observation of our corpus, we can say that there

is no correlation with prosodic focalization: reinforcing gestures are not associated with any specific stress type. In the same way, there is no correlation with eyebrow movements. On the contrary, there is a clear link with adverbs and connectors at the beginning of speech turns. In fact, reinforcing gestures plays an important role in discourse planning; they play a syntactic role more than a focusing one. They are in conclusion more discursive than expressive.

6. Conclusion

Annotated multimodal corpora constitute an essential resource in modern linguistics. The understanding of language mechanisms (both in production and perception) needs to take into account very precisely the interaction between all the different domains or modalities (phonetics, prosody, lexicon, syntax, pragmatics, gestures, etc.). Producing such resources represents however a huge amount of work. It is then necessary to specify a precise framework, identifying the different tasks, the kind of information they have to produce and to what extend they can be automatized. We have presented in this paper an experiment exploiting different tools and shown how they can be integrated in order to provide a platform.

7. References

Allwood J., L. Cerrato, L. Dybkjaer, & al. (2005). The MUMIN Multimodal Coding Scheme, NorFA yearbook 2005, http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf

Bertrand, R., Blache, P., Espesser, R., & al., « Le CID: Bertrand, R., Blache, P., Espesser, R., & al., (2007a) « Le CID: Corpus of Interactional Data – protocoles, conventions, annotations », *TIPA*, 25:25-55,.

Bertrand, R., G. Ferré, P. Blache, R. Espesser, and S. Rauzy (2007b) "Backchannels revisited from a multimodal perspective." *Proceedings of Auditory-visual Speech Processing,* Hilvarenbeek, The Netherlands., Cederom.

Blanche-Benveniste, C & Jeanjean, C. (1987). Le français parlé, Transcription et édition. Paris: Didier-Erudition/ InaLF, 2e éd.

Carletta J. & Isard A. (1999), The MATE Annotation Workbench: User Requirements, in Proceedings of the ACL Workshop: Towards Standards and Tools for Discourse Tagging, pages 11-17, University of Maryland, June 1999.

Di Cristo, A. & Di Cristo, P. (2001). Syntaix, une approche métrique-autosegmentale de la prosodie. TAL, 42(1), p. 69-111.

Di Cristo A., Auran C., Bertrand R.**,** et al., (2004). Outils prosodiques et analyse du discours », in A.C. Simon, A. Auchlin et A. Grobet (eds), *Cahiers de Linguistique de Louvain 30/1-3,* Louvain-la-neuve : Peeters, 28, p. 27-84.

Ferré, G., R. Bertrand, P. Blache, R. Espesser, and S. Rauzy. (2007) "Intensive Gestures in French and their Multimodal Correlates." *Proceedings of Interspeech,* Antwerp, Belgium), Cederom.

Hirst, D., Di Cristo, A. & Espesser, R. (2000). Prosody : Theory and Experiment, chapter Levels of description and levels of representation in the analysis of intonation. Kluwer : Dordrecht, Pays-Bas. p. 51-87.

Isard A. (2001). An Xml Architecture For The Hcrc Map Task Corpus. In P. Kuehnlein, H. Rieser, H. Zeevat, Eds, Bi-Dialog 2001.

Kendon A. (2004). Gesture : Visible Action As Utterance. Cambridge: CUP.

Kipp M. (2004). Gesture Generation By Imitation. From Human Behavior To Computer Character Animation. Florida, Boca Raton (http://www.dfki.de/~Kipp/Dissertation.html)

Krenn B. & Pirker H. (2004). Defining The Gesticon: Language And Gesture Coordination For Interacting Embodied Agents. Aisb-2004 Symposium On Language, Speech And Gesture For Expressive Characters, University Of Leeds, UK.

McNeill D. (2005) Gesture and Thought. Chicago: University of Chicago Press.

Tusnelda (2005). Tübingen collection of reusable, empirical, linguistic data structures. http://www.sfb441.uni-tuebingen.de/tusnelda-engl.html

Van Rullen, T., "Vers une analyse syntaxique à granularité variable", PhD Thesis, University of Aix-Marseille I.